http://lexikos.journals.ac.za

How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study

Robert Lew, *Department of Lexicography and Lexicology, Faculty of English, Adam Mickiewicz University in Poznań, Poland (rlew@amu.edu.pl)*

Marcin Grzelak *(grzelak.m.r@gmail.com)*
and
Mateusz Leszkowicz, *Department of Pedeutology, Faculty of Educational Studies, Adam Mickiewicz University in Poznań, Poland (mateuszl@amu.edu.pl)*

Abstract: We use modern eye-tracking technology to scrutinize the process of sense and equivalent selection in polysemous bilingual entries. Our study subjects, intermediate and advanced Polish learners of English, consulted 26 Polish-to-English dictionary pages prompted with a sentence translation task. Throughout the task, an eye-tracking device unobtrusively recorded their gaze patterns, which are analyzed and discussed. Both successful and unsuccessful searches are examined. Also, we assess the potential of eye-tracking technology in the study of dictionary use.

Keywords: EYE TRACKING, BILINGUAL DICTIONARY, ENTRY NAVIGATION, INNER ACCESS, MICROSTRUCTURE, POLYSEMOUS ENTRIES, SENSE INDICATION, ENGLISH LANGUAGE, POLISH LEARNERS

Opsomming: Hoe woordeboekgebruikers betekenisse in tweetaligewoordeboekinskrywings kies: 'n oogvolgstudie. Ons gebruik moderne oogvolgtegnologie om die proses van betekenis- en ekwivalensiekeuse in poliseme tweetalige inskrywings te ondersoek. Ons studiepersone, intermediêre en gevorderde Poolse aanleerders van Engels, het 26 Poolsna-Engelse woordeboekbladsye geraadpleeg vir die doel van 'n sinsvertalingstaak. Gedurende die hele taak het 'n oogvolgtoestel onopsigtelik hul kykpatrone geregistreer wat ontleed en bespreek word. Sowel suksesvolle as onsuksesvolle soektogte word beskou. Ook beoordeel ons die potensiaal van oogvolgtegnologie by die studie van woordeboekgebruik.

Sleutelwoorde: OOGVOLGING, TWEETALIGE WOORDEBOEK, INSKRYWINGNAVIGASIE, BINNETOEGANG, MIKROSTRUKTUUR, POLISEME INSKRYWINGS, BETEKENISAANDUIDING, ENGELSE TAAL, POOLSE AANLEERDERS

1. Introduction

Researchers working in the area of dictionary use have always wanted to be able to observe which parts of dictionary entries dictionary users are paying attention to when engaging in dictionary consultation, and in what sequence. In search of clues as to which particular portions of an entry are used and when, investigators have resorted to introspective techniques based on self reports. These might variously involve underlining, think-aloud protocols, or written self-recording sheets.

In studies which employ underlining (e.g. Bogaards 1998; Lew and Dziemianko 2006; Lew 2010), users are usually asked to physically mark in pencil the particular fragments of the dictionary entry which they happen to be consulting while engaged in a task calling for lexicographic support. There are several quite serious downsides to this technique. One is due to attentional factors: asking someone to underline text in a dictionary ties up a portion of their attention, which is otherwise busy with both the dictionary consultation task and the primary task for which dictionary assistance is sought. Another problematic issue is the degree to which the effort of monitoring which parts of a dictionary are being used affects the very way in which dictionary users make use of a dictionary. It is quite likely that the distortion is substantial, as they need to be constantly aware of the monitoring aspect, as they attend to registering details of their dictionary consultations. Finally, there is a real danger of some of the consultation activity being left unrecorded: participants fail to underline as expected, as they become focused on the main task.

Similar problems beset the use of think-aloud protocols for recording dictionary consultation (Al-Besbasi 1991; Mackintosh 1995; Wingate 2002). Here again we are likely to obtain a self-conscious, incomplete, and distorted picture of consultation behaviour. The degree of success with this technique depends substantially on the skills of the participants in following the protocol.

Written protocol sheets are another option (e.g. Harvey and Yuill 1997). But if completed during the consultation, they tend to be even more intrusive and distracting than underlining. As their complexity naturally reflects the structural involvedness of dictionaries themselves, the quality of the data returned is questionable. In contrast, retrospective protocols completed after dictionary consultation will fail to record the interesting detail due to memory limitations.

All of the above options suffer from problems which severely diminish the validity of recording dictionary consultation in any of these ways. Until very recently, there had been no way to collect reliable information on which sections of the dictionary entry the user was consulting and in what sequence. But such an option became a possibility with the advent of modern eye-tracking technology. Today's equipment allows non-intrusive monitoring of participants' gaze, yielding insights into the patterns of dictionary consultation.

This study uses eye tracking to look into the process of inner access, that is

entry-internal navigation. Our participants were instructed to look for the sense which held an English translation equivalent appropriate for the context of the Polish sentence cue presented with the entry. As eye tracking has been used very little in dictionary user studies so far, another goal of the study is to examine the applicability of this technique to the study of dictionary entry navigation.

The design of the study will be explained in section 4, and its results will be presented in section 5. Before we get to the study itself, however, some background on eye movement research will be given below (section 2), focusing on its application within dictionary user studies, followed in section 3 by an overview of previous studies dealing with sense selection. As we want to leave as much space as possible to presenting the results of the study, we try to keep the overview sections brief. Readers wishing to learn more about eye tracking are invited to consult Rayner (1998), whereas Nesi and Tan (2011) offer a comprehensive overview of findings on sense selection.

2. Eye tracking in dictionary user research

Eye tracking, also known as gaze tracking or eye movement recording (EMR), is by no means a new technique: it has been in sporadic use for over 100 years now. But it was only recently that advances in technology made it an attractive and affordable option for researchers in many domains, most importantly in the examination of various aspects of reading and visual processing, human–computer interaction and web design.

Eye tracking covers a cluster of related techniques for monitoring and recording fine movements of the eye as an indication of where the subject is looking at a given time, as well as the sequencing of gaze movements across some visual-perceptual space.

Human visual perception normally involves a series of intermittent *fixations*, during which the location of the gaze is relatively stable, and *saccades*, which are quick movements in between the consecutive fixations, when no significant visual processing takes place. Gaze behaviour is usually interpreted as reflecting perception, on the strength of the *eye–mind assumption* (Just and Carpenter 1980). A fixation then is assumed to represent perceptual and cognitive processing of stimuli. In reading a text, longer fixations imply longer processing, possibly due to increased attention, such as when facing some difficulty. Gaze *regressions* in reading are movements against the normal text orientation: in English, regressions are movements to the left, as in any left-to-right writing language. More extensive regressions may take the gaze back to a previous line. These are often indicative of the reader backtracking during reading in order to resolve a processing problem. Parameters of eye movement such as gaze duration, saccade length or search time are believed to correlate well with processing complexity (Duchowski 2007). Eye fixations reflect the encoding of

information based on the stimuli being viewed. Research on reading, mostly English texts by fluent native speakers, typically finds mean fixation duration for single fixations to be on the order of 225 milliseconds for silent reading and 275 milliseconds for reading aloud, although specific values vary both individually and with text difficulty (Rayner 1998: 373, 2009). There are no reliable data as yet on eye movement parameters during the reading of a dictionary entry, which is a formally structured special text, qualitatively different from normal reading matter.

Only a handful of studies of dictionary use have so far employed eye-tracking technology. Three of these studies are due to Henrik Køhler Simonsen. Simonsen (2009a) investigated gaze patterns and gaze duration of users consulting an online Danish accounting dictionary, looking for evidence of differences in reference behaviour associated with different lexicographic functions, or modes of using the dictionary (L1 knowledge acquisition, L1 production, L1 reception, and L1-to-L2 translation).

In the same year, Simonsen (2009b) compared gaze parameters of professional translators working with vertical and horizontal data presentations in an internet dictionary. He was also concerned with general methodological issues of viability of eye tracking for studying internet dictionary consultation. In Simonsen (2011), the author further explored the applicability of the analysis options typically present in eye movement data analysis software for dictionary user reference behaviour.

A study by Kaneta (2011) looked at the frequency and duration of reference to illustrative examples in two forms of digital entry presentations: unfolded (flat) and folded (layered), in both monolingual and bilingual entries. Rather predictably, when illustrative examples were hidden from the initial view, they were consulted less often than in a complete presentation.

Tono (2011) used eye-tracking technology to examine a number of variables related to look-up behaviour. This work is especially relevant to the present study, as his investigation focused on entry navigation devices (menus and signposts). It will be summarized in the following section.

3. Previous studies of sense navigation

Tono (1984) was a pioneering work addressing the issue of how users select senses in entries. In this study, dictionary users exhibited a tendency to pick the first sense of a bilingual entry and ignore the remainder of the entry, unless the first sense did not fit in an obvious way. A number of subsequent studies focused on entry navigation devices in the form of (entry-initial) menus and (sense-initial) signposts, typically in monolingual entries (Tono 1992, 1997; Bogaards 1998; Tono 2001; Lew and Pajkowska 2007; Lew 2010; Nesi and Tan 2011; Tono 2011), though not exclusively (Lew and Tokarek 2010). Most studies have confirmed the value of signposts, both in terms of helping users find the

right sense, and in terms of speed. Signposts were found to be more effective than menus by Lew (2010) and Nesi and Tan (2011), but Tono (2011) reports the opposite.

An important challenge to previous findings on the advantage of entry-initial senses came from Nesi and Tan (2011), who found entry-*final* senses to be at least as salient as entry-initial ones. These disparate results are not necessarily contradictory if we allow for the fact that entry navigation strategies may be contingent on several factors, including the user's reference skills, proficiency in language, type and form of the dictionary, properties of the lexical item being looked up, and the task which prompted dictionary consultation. It is quite possible that the final-sense advantage arose in this case as part of a specific consultation strategy of fairly experienced dictionary users consulting English monolingual learners' dictionaries in comprehension tasks. Such users may have discovered through continued dictionary work that the most relevant senses are usually found towards the bottom of polysemous entries, as the most frequent senses at the top are usually familiar.

There are very few studies investigating sense identification in bilingual dictionaries (Lew and Tokarek 2010). Sense guidance in L2→L1 bilingual dictionaries is largely achieved by virtue of the fact that entries feature equivalents in the users' native language. Thanks to their salience in the respective senses, such entries can usually be scanned quite efficiently, and the need for additional navigation aids is diminished. In contrast, L1→L2 dictionaries feature equivalents in a language of which the user has only partial knowledge. Reading, and especially scanning, foreign language text is obviously less efficient than in one's native language. Also, many of the L2 equivalents given in an entry will not be well known to the dictionary user, and thus provide few clues to meaning. Multi-word expressions and phrases in the source language (L1) may offer useful visual pivots in those entries that cover them, but in order to distinguish between decontextualized equivalents, the better bilingual dictionaries supply sublemmatic guiding elements in the form of sense indicators and equivalent discriminators. Sense indicators in bilingual dictionaries and signposts in monolingual entries are in fact quite close, both structurally and functionally.

A study by Tono (2011) deserves special attention in this context as it is similar to the present study in both its goals and use of an eye-tracking system. Tono attempted to test several variables at a time, and the results are somewhat complex and difficult to interpret unambiguously. Perhaps the most important finding to take out of Tono (2011) is that consultation behaviour is rarely systematic, but tends to be erratic. This suggests that a neat, simple model of dictionary consultation which implicitly underlies many lexicographers' efforts may be too much of an idealization. Tono concludes by calling for further study of the look-up process with the use of eye-tracking systems. The present paper responds to this call.

4. The study

4.1 Aim

The overall aim of this study is to examine how users navigate polysemous bilingual dictionary (L1 to L2) entries in a lexical search scenario induced by a sentence translation task from L1 (Polish) to L2 (English). Further goals are:

— to uncover patterns of dictionary users scanning to locate senses in a bilingual dictionary, as revealed by eye-tracking data;

— to compare successful and unsuccessful searches and explore their possible correlates; and

— to explore the applicability of the eye-movement paradigm to the study of dictionary entry navigation.

4.2 Participants

Participants in the study were ten Polish university students. Half of them were English majors and thus advanced learners of English (CEFR level B2 to C1). These participants made up the high-proficiency (HP) group, and reported using dictionaries on a daily basis. The other five participants were majoring in the following areas: preschool education; modern Greek; architecture and town planning; environmental engineering; and corporate management. These five low-proficiency (LP) students (CEFR level A2 to B1) admitted to using dictionaries several times a week. All recruited subjects had full visual acuity (20/20), some with correction. The eye-tracking system we used is tolerant of correction glasses and (untinted) contact lenses, so this was not an issue.

4.3 Materials

Two sets of polysemous Polish-to-English dictionary entries were adapted from two modern Polish–English bilingual dictionaries: thirteen items from *Nowy Słownik Fundacji Kościuszkowskiej. The New Kosciuszko Foundation Dictionary* (NKFD 2003) and another thirteen items from *Wielki Słownik Angielsko–Polski, Polsko–Angielski PWN-Oxford* (PWNO 2002). We chose these two dictionaries as they are currently the only comprehensive general bilingual dictionaries between Polish and English compiled and published in this century. They are comparable in size and coverage. Both dictionaries use similar techniques and devices for sense guidance, with Polish as the metalanguage in the respective Polish–English volumes. Senses are most typically indicated by means of near-synonyms, hyponyms, collocates or domain labels, which is standard fare in general bilingual dictionaries of high quality. Dictionary page mock-ups were constructed for the twenty-six items, replicating the original typography

closely, but increasing the font size and line spacing so as to make the text comfortable to read on screen from a viewing distance of 60 cm by subjects with normal vision. Each word was presented in a context sentence constructed for the experiment, and placed in a rubric above the dictionary extract which included the entry for the target word. Experimental stimuli were prepared as single-page PNG graphic files at a native Tobii T60 screen resolution (1280 x 1024 pixels) so as to avoid interpolation and aliasing distortion. The top 9% (90 pixels) of the screen included a frame with the sentence cue against a light-grey background. The remainder of the screen held the mock-up dictionary page with the test entry. Shorter entries were accompanied by alphabetically neighbouring entries, as they would in a paper dictionary. A typical stimulus item with a dictionary page mock-up and a sentence cue appeared as in **Figure 1**. There were twenty-six screen pages, each containing a sentence cue with an underlined key word, including thirteen entries from each dictionary.

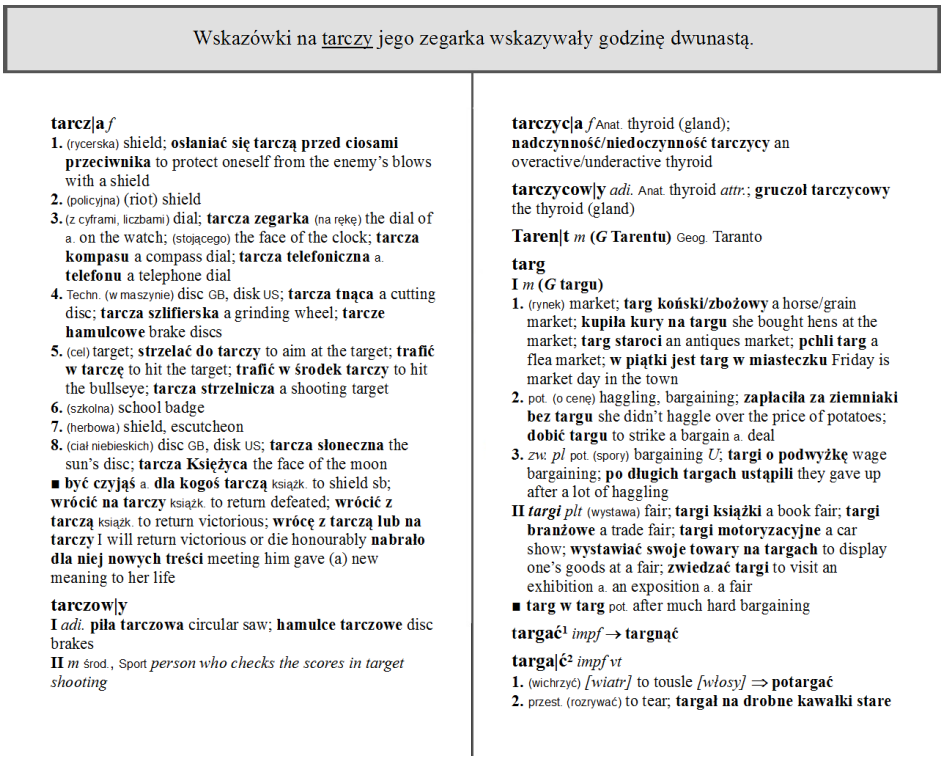

Figure 1: A sample stimulus item with a dictionary page.

The twenty-six Polish key words were: kosz, siatka, poślizg, blok, ekspozycja, emisja, język, forma, przedmiot, siła, rakieta, rezerwa, promień, paczka, treść, płyta, album, dyscyplina, figura, legenda, korek, prąd, wpaść, serce, tarcza,

podziałać. The criteria guiding the selection of words and target senses were: familiarity, difficulty, and sense position in the dictionary. We chose familiar Polish words, but in less frequent senses which, though clear to the participants at both levels, would present a challenge in selecting their English equivalents. We wanted participants to focus on picking the correct dictionary sense with its English equivalent rather than puzzle over the meaning of the Polish word. For example, the familiar Polish word *poślizg* was used, not in its default sense 'skid', but in the metaphorically derived sense 'delay'. The context of each sentence cue made it clear which sense was meant, which was verified during the piloting stage.

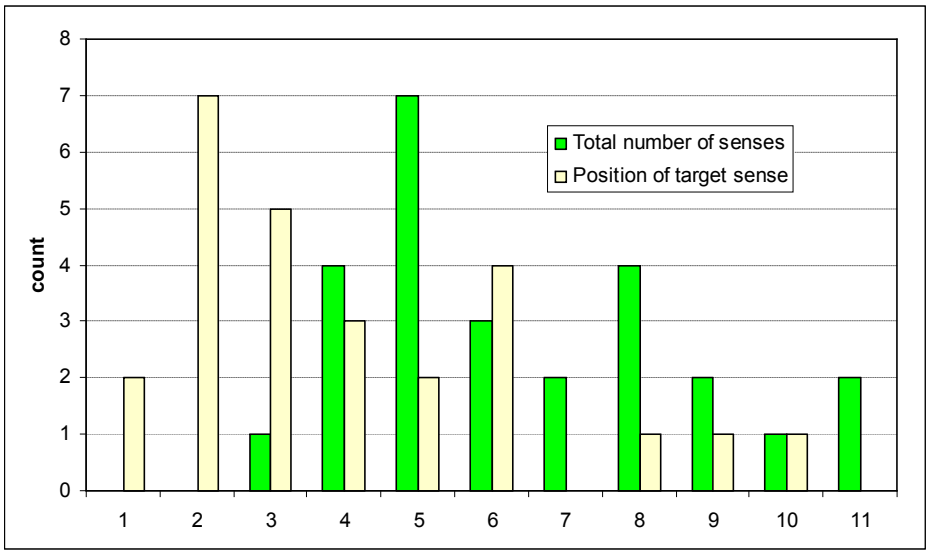

Figure 2: Distribution of the (1) total number of senses in the entry and (2) position of the target sense within the entry.

We have included entries of between three and eleven senses in length, whereas the position of the target sense in the entry ranged between first and tenth. A detailed distribution of the total number of senses in the entry and the position of the target sense is given in **Figure 2**. For example, at x=5, the two differently-shaded bars indicate, respectively, that the complete set of 26 items included seven entries of five senses, and there were two entries in which sense five was the target sense. Mean length of entry was 6.5 senses (median = 6). Mean position of target sense was 4.0 (median = 3). Although it was not our intention to compare the two dictionaries, we nevertheless made an effort to balance entry length and target sense position across the dictionaries. Thus, for PWNO and NKFD, respectively, mean length of entry was 6.4 (median = 7) and

6.5 (median = 5), and mean position of target sense was 3.8 (median = 3) and 4.2 (median = 3).

4.4 Apparatus

A Tobii T60 eye tracker was used in the study. The unit has a sampling rate of 60 Hz, a nominal spatial accuracy of at least 0.5 degrees and drift below 0.3 degrees of visual angle. It is equipped with a 17-inch screen with a native resolution of 1280 x 1024 pixels. An important advantage of this model is that it requires no head restraining, allowing the participant a fair amount of head movement. The tracking cameras are integrated in the main unit and are not obvious to the user. The unit looks like an ordinary flat-screen computer monitor. Thanks to these features, the Tobii T60 has high ecological validity, offering participants the look and feel of a regular computer screen, thus a highly naturalistic setting for students accustomed to working with a computer. The software used in the design of the experiment and data collection during the recording sessions was Tobii Studio, version 2.0.8.

4.5 Procedure

The experiment took place in a spacious, daylit university office. A single experimenter (the second author) worked individually with one participant at a time, in several sittings during January 2012. Participants would be seated in front of the Tobii T60 unit placed on top of a work desk, at a viewing distance of about 55 centimetres from the monitor, with the eyes at a level just below the centre of the screen. The unit was connected to two computers: one running the Tobii Studio software to control the tracking unit and collect the data, the other for the researcher to monitor progress. During the experiment, the experimenter was able to monitor the participants' posture on his screen, and correct it if needed.

Before the recording session, each participant was given specific instructions (in Polish) about the procedure. Participants were asked to keep looking at the screen and try not to move their heads too much. Then the eye tracker was calibrated, once for each participant. The participants followed a red dot on the screen with their eyes as it moved around for a few seconds. Once the calibration was successful, the recording could begin.

Each participant was presented with the same twenty-eight screens in turn. The first screen included basic instructions reminding the participants what they should focus on. The last screen indicated the end of the recording. The other twenty-six screens included dictionary pages with sentence cues as described above. The order in which the twenty-six experimental items appeared was randomized to minimize any order effects. For each of the twenty-six items, participants were asked to translate the underlined word by locating the appropriate sense within a polysemous entry presented below the

context sentence. They were asked to speak the translation out loud after they had decided on the correct equivalent. A complete audio transcript of the sessions was made using a digital recorder. We had rejected the option of asking the participants to write down the answers themselves, as this would have made them look away from the monitor and might have disrupted the gaze recording. We did not want to ask them to give the sense number itself, as this might have made them too aware of the sense selection aspect. With the solution adopted, the equivalent itself did not in every case unambiguously indicate which particular sense was chosen, as sometimes different senses shared an equivalent. However, coupled with a subsequent review of eye-scan paths, the sense selected could always be determined with high certainty. The complete procedure was piloted on two students to ensure that all elements worked as expected. The main experiment proceeded smoothly with no problems, yielding good quality eye-tracking data, which were subsequently analyzed. The only slight complication was that the unit we used exhibited spatial accuracy issues for the extreme upper area of the screen. In recording such extreme top values, it tended to offshoot towards the margin of the screen in the vertical dimension. The problem was caught during initial testing (even before the piloting), and so the top strip of the screen was used for the sentence cue. Since twenty-three of the twenty-six sentence cues fit on a single line of text, this did not cause any ambiguity in interpreting the data, and in any case our main interest was not in how participants read the sentence cue, but how they worked with the dictionary excerpt below. Importantly, in the entire area of the screen used for the dictionary mock-up there were no spatial accuracy issues, so analysis by Areas of Interest could proceed without distortion.

4.6 Data analysis

The experiment generated 260 complex searches (data from ten participants, each looking up twenty-six items). Each search was classified as successful or unsuccessful, depending on whether the participant located the contextually correct sense in the correct entry. For a successful search, two conditions had to obtain at the same time: (1) the participant had to provide the correct English equivalent for the sentence cue in the verbal feedback; and (2) an examination of the gaze paths showed fixations on the target sense which coincided with the correct English equivalent.

Complete eye movement data from all 260 searches were collected. Measures used in the analysis included fixation counts and fixation duration. To detect fixations, the Tobii sliding-average algorithm was used as described in Olsson (2007), with the default threshold radius of 35 pixels. This filter setting turned out to be very effective in detecting fixations in both the inner access searches (within the entry) and outer access searches (headword scan). We also tested the ClearView filter at the settings recommended by Gerganov (2007) for translation-related data: 80 milliseconds minimum fixation duration, and 40

pixels fixation radius. This setting worked reasonably well for inner access searches, but missed some of the quicker headword scans which were apparent in the raw data. In order to capture those as well, the ClearView filter had to be set at 40 ms for minimum fixation duration and at least 60 pixels fixation radius. Visual scan paths were generated for all searches to aid in the qualitative assessment of look-up behaviour.

Areas of Interest (AOIs) were defined and plotted manually for all twenty-six dictionary excerpts. AOIs were entered separately for each entry so that in each case they covered the following entry components (refer to **Figure 3**):

— the headword (coded as *hw*);

— each sense (coded as *sN*, where *N* was the relevant sense number); and

— sense-guiding element(s) for each sense and any embedded subsenses or phrases within the sense (coded as *gN*, where *N* was the relevant sense number).

This was done in order to allow the computation of gaze data specific to the structural components of the entries: individual senses and their guiding elements.

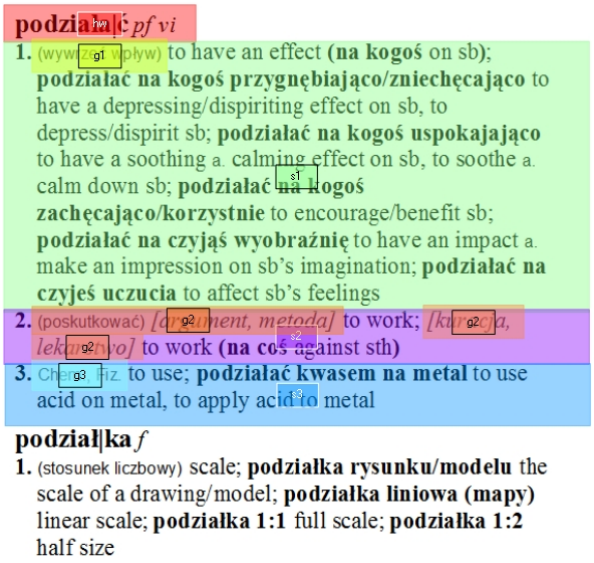

Figure 3: Areas of Interest (AOIs) marking and coding (hw = headword; s*N* = sense number *N*; g*N* = guiding element for sense *N*)

Fixation counts and fixation duration were computed for the above entry components, by participants and by items separately. In addition, separate calculations were made for the structural elements of the target sense: the target sense

itself and any of its guiding elements.

The eye tracker also logged the time stamp for every event, so accurate timing data became available as well, and these were used in calculating temporal parameters of the consultations.

The software used in the analysis included: Tobii Studio 2.0.8, OpenOffice Calc 3.4.1, LibreOffice Calc 4.0.0.3, Microsoft Excel 2003, Statistica 8.0, and statistical test applets at http://www.vassarstats.net (Lowry 2001–2013).

5. Results and discussion

5.1 Task completion time

The mean time for a single participant to complete the whole task (all 26 items) was 557 seconds. Considering the two groups of participants separately, an average high-proficiency student needed 509 seconds to finish the task, while in the low-proficiency group the mean time spent on the task was 19 per cent longer at 605 seconds. This difference, however, is not statistically significant (one-way ANOVA, $F_{(1, 8)} = 0.94$, $p = 0.36$, partial $\eta^2 = 0.11$), likely due to a small number of subjects coupled with large individual variation. A detailed breakdown of the time taken to complete the whole task is given in **Table 1**.

Table 1: Task completion times for individual participants in the two groups

low proficiency		high proficiency	
participant	total time [sec]	participant	total time [sec]
LP-03	652	HP-01	713
LP-06	615	HP-02	500
LP-07	844	HP-04	346
LP-08	514	HP-05	591
LP-09	401	HP-10	394
LP mean	605	HP mean	509
LP SD	165	HP SD	149

Next, task completion times were computed for each item (see **Table 2**). An average item took 20.1 seconds to complete (*SD* = 6.6 seconds). Mean values computed for NKFD and PWNO items separately were 19.1 and 21.1 seconds, respectively. This difference is not statistically significant (one-way ANOVA, $F_{(1, 24)} = 0.61$, $p = 0.44$) and its effect size is very weak (partial $\eta^2 = 0.02$). In addition, it might be noted that one of the PWNO items (*płyta*) took more than twice the average time to finish, and this item is responsible for the high standard deviation value in the PWNO set. If this one outlier is ignored, the mean for PWNO goes down to 19.2, and becomes virtually identical to the corresponding NKFD figure.

Table 2: Task completion times for individual items and the two dictionaries

NKFD		PWNO	
item	mean time [sec]	item	mean time [sec]
blok	16.8	album	13.1
ekspozycja	20.9	dyscyplina	15.7
emisja	16.2	figura	14.4
forma	23.8	korek	17.0
język	18.9	legenda	13.4
kosz	21.6	paczka	18.5
poślizg	17.9	płyta	43.6
promień	17.0	podziałać	28.5
przedmiot	18.5	prąd	14.6
rakieta	12.8	serce	24.8
rezerwa	14.9	tarcza	21.4
siatka	24.6	treść	20.5
siła	23.9	wpaść	28.8
NKFD mean	19.1	PWNO mean	21.1
NKFD SD	3.67	PWNO SD	8.67

5.2 Sense selection success rates

There were 260 searches overall (26 items x 10 participants). Of those searches, 157 covered all senses in the entry, i.e. there was evidence in the eye-movement data of participants fixating on every sense within an entry. The remaining 103 searches failed to examine one or more senses within an entry. Nineteen percent of all searches (50 out of 260) resulted in the wrong sense being chosen. One might hypothesize that errors in sense selection might be more likely in incomplete searches, and less likely in searches covering all senses, as incomplete searches carry a greater risk of missing the target sense. To assess whether this was actually the case, we tabulated error counts separately for complete and incomplete searches in **Table 3**. Altogether, there were 50 searches resulting in errors of sense selection. Of these, 26 errors occurred in complete searches (an error rate of 17%, or a success rate of 83%), and 24 errors in incomplete searches (an error rate of 23%, success rate 77%). The difference between the proportions of errors in the two types of searches in not statistically significant (Z-test for independent proportions, z-score = –1.349, p = 0.18). The error rates were thus quite similar across the two types of searches, suggesting that viewing all senses was no guarantee of getting the sense right. Conversely, incomplete searches could well be successful: these searches typically stopped once the target sense was positively identified.

Table 3: Proportion of searches resulting in erroneous sense selection, when all senses have been examined, and when some senses have been omitted

	all searches	wrong sense selected	correct sense selected	sense selection error rate	sense selection success rate
all senses viewed	157	26	131	17%	83%
not all senses viewed	103	24	79	23%	77%
all searches	260	50	190	19%	81%

5.3 Sense selection success, entry length, and sense position

To investigate the relationship between sense selection success and entry length, we computed separate Pearson correlation coefficients between sense selection success rates and three measures: (1) total number of senses; (2) absolute target sense position; and (3) relative target sense position within the entry. None of these correlations turned out to be statistically significant, and detailed results are given in **Table 4**.

As shown in the top row of **Table 4**, there is a weak positive correlation (r = 0.28) between the total number of senses in the entry and sense selection success rate. Therefore, there is no evidence here that it was easier to find the correct sense in shorter entries than in entries with a larger number of senses. The tendency is actually the reverse, although it is not significant (p = 0.16).

Table 4: Pearson correlations (r) between the sense selection success rate and (1) total number of senses; (2) absolute target sense position; and (3) relative target sense position. Also given are: coefficient of determination (r^2), and the t-score and p-level for H_1: $r \neq 0$ against H_0: $r = 0$)

	r	r^2	t	p
total number of senses	0.28	0.08	1.4	0.16
absolute target sense position	−0.03	0.001	−0.13	0.90
relative target sense position	− 0.29	0.08	−1.5	0.16

The second row of **Table 4** reveals a near-zero correlation, or an almost perfect independence, between the success rate and sense number of the target sense. This suggests that the ease of locating the target sense was independent of how

far from the beginning of the entry the relevant sense was located.

In the last row, correlation is given of the success rate with a relative measure of target sense position. This was computed by dividing the target sense number by the total number of senses in the entry. For example, sense number five in an entry of ten senses would have a relative position of 5/10 = 0.5 (or 50%). This measure exhibits a small degree of negative correlation with the success rate of sense selection. There is thus a slight tendency for target senses that are relatively early in the longer entry to be somewhat easier to find, but this tendency is not statistically significant ($p = 0.16$). This best estimate of the correlation coefficient yields a low coefficient of determination of $r^2 = 0.08$; this means that only about 8% of the success in locating the target sense could be accounted for by the relative placement of the target sense within the entry.

Remarkably, there was a perfect match in all searches between target sense selection success and correct provision of the contextually appropriate English equivalent. This means that once participants were able to locate the correct sense, they had no difficulty extracting the right equivalent. In most cases, high-proficiency participants were able to locate the relevant sense, pick the correct English equivalent and use it in the translation of the sentence cue. Of the five high-proficiency participants, two did not make any sense selection errors at all, one made a single error, another one committed two errors, and the one least successful HP participant made five errors in the 26 items. This translates into an overall error rate of six percent only in the HP group. Selection errors were far more frequent in the low-proficiency group, with 42 erroneous sense selections out of a total of 130 individual lookups, that is 32% of all cases, or over five times the error rate in the high-proficiency group. The errors were more or less evenly distributed across the individual low-proficiency students.

Interestingly, nearly all cases (with only two exceptions) of incorrect selection involved choosing a sense earlier than the target sense. Specifically, in sixteen cases of erroneous sense selections, the sense selected was placed one sense above the target item, and in nineteen cases (the most typical scenario) a sense preceding the target by two senses was chosen. More tellingly, though, of the 50 erroneous selections, 35 (or 70%) involved the participants' choosing the first sense of an entry. This corroborates the results of some previous studies (Tono 1984; Lew 2004), where users were found to pick the first sense if they did not know any better. In contrast, we found no evidence in our data of an advantage of entry-*final* senses which was reported by Nesi and Tan (2011). In fact, not a single erroneous choice in our study involved a participant wrongly selecting an entry-final sense. Further, detailed eye-tracking data revealed that users typically proceeded from the top of the entry downwards rather than from the bottom up. This possible difference in reference behaviour may find explanation in the fact that our study examined bilingual entries, unlike in Nesi and Tan (2011), where entries from English monolingual learners' dictionaries

were used. These entries tended to be arranged by sense frequency, with the rarest senses appearing towards the end, and an awareness of this organizing principle might have prompted Nesi and Tan's subjects to work from the final senses upwards. But bilingual dictionaries tend to be organized differently, with the range of applicability of the translation equivalent playing an important role, and in this case the traditional top-down scanning of entries as predominantly used by our participants may represent the better default strategy.

Of the total fifty misidentified senses, eleven errors were made in WSPO, and thirty-nine in NKFD. This might invite the conclusion that WSPO is more user-friendly in its sense guidance. However, we should keep in mind that the experiment was not designed to compare the two dictionaries. In contrast to Tono (2011), we preferred not to present the same headword more than once to the same participant, mindful of the risk of carry-over effects. In our experiment, the two sets of entries from the two dictionaries (thirteen each) had all different headwords. Although we made an effort to match the two sets in terms of sense position and entry length (see section 4.3 above), there are often subtle and unpredictable effects in language data which make some items more difficult to the experimental subjects than others. It may be that a greater number of such difficult items found their way into the NKFD set, negatively impacting consultation success in these cases. Another possible confounding factor is the familiarity of participants with the dictionaries. WSPO is probably the more popular of the two modern comprehensive Polish–English dictionaries, and thus more likely to be known. Having said this, it must be stressed that the two dictionaries are fairly similar in terms of sense discrimination strategies, layout and typography, which makes item difficulty the more probable reason for the observed difference.

5.4 Gaze fixation statistics

Overall mean duration of fixation was 298 milliseconds (SD = 190 ms): that is only about 30 per cent longer than the typical value for silent text reading by native speakers (Rayner 1998). This would suggest that, in terms of eye-movement parameters, consulting dictionary entries is not dramatically different from normal text reading, and experience gained in extensive eye-movement research in reading can, with some caution, be drawn on in dictionary user studies. Fixation duration was fairly uniform across our ten participants. Mean per-subject values ranged from 262 ms (participant HP-05) to a high of 319 ms (participant HP-02). Means calculated for each of the groups separately yield 307 ms for low-proficiency and 290 ms for high-proficiency participants, which is a modest difference and not statistically significant (one-way ANOVA, $F_{(1, 8)}$ = 2.05, p = 0.19, partial η^2 = 0.20).

Across all items, participants made 9267 fixations within any of the senses. Of these, 22% (2019) were focused on sense guiding elements. This proportion was slightly higher for high-proficiency participants (23% on average) than for

low-proficiency participants (21%). This small difference turned out not to be statistically significant (one-way ANOVA, $F_{(1, 8)} = 0.43$, $p = 0.53$, partial $\eta^2 = 0.05$).

In terms of total *dwell time* (i.e. cumulative duration), fixations on guidewords accounted for 23% of the time spent looking anywhere within any of the senses. Here again, a slightly higher proportion of time spent on guiding elements is evidenced in searches by high-proficiency participants (25% of the time on average) than those by low-proficiency participants (21% of the time). Similarly as in the case of fixation counts, this difference was not statistically significant (one-way ANOVA, $F_{(1, 8)} = 1.12$, $p = 0.32$, partial $\eta^2 = 0.12$).

The above figures indicate that there was a tendency, albeit not statistically significant, for lower-level participants to fixate for longer periods. With regard to the proportion of attention directed towards sense guiding elements, the two proficiency levels used guiding devices to a similar extent, unlike in Tono's (2011) study. This suggests that sense guidance was universally useful.

5.5 Patterns of look-up behaviour

The main aim of this study has been to examine how dictionary users look up senses in polysemous bilingual entries in translation-induced production, and how the position of the sense affects the process. This major section deals with this issue with the help of two types of visualizations of eye-tracking data: scan paths and heat maps.

A systematic qualitative analysis of the scan paths of all the lookups revealed that by far the dominant strategy was to engage in a systematic scan of the senses, starting at the top of the entry and proceeding in a downwards direction until the last sense was reached. Participants usually scanned rather rapidly through the senses, mostly focusing on their sense indicators, until they reached what they believed was the right sense. At that point, they would normally proceed to read the entire sense rather more carefully. A typical example of such a pattern of consultation behaviour is mapped out in **Figure 4**. In this scan path representation, fixations are shown by dots (the larger the dot, the longer the duration). The numbers in the dots represent their temporal sequencing.

Participants would normally start by reading the sentence cue in Polish (fixations 3-6). The first brief fixation or two would sometimes be elsewhere (just as the new stimulus first appeared), not uncommonly within the central area of the screen (this is normal). Once in the sentence cue, they would soon focus on the underlined word, whose equivalent was sought (here the longer fixation number 7). Next, they would scan the headwords in search of the lemma sign. In this particular case, as is usual in a highly-inflected language such as Polish, they would need to reduce the inflected noun form (*formie*, locative) to its citation form (*forma*, 'form'). Here the scanning did not start at the top left of the dictionary page, where the lemma sign *forma* was actually to be found, but rather in a vertical motion down to the headword *formacja* (fixa-

tion 9). This was not the right lemma, but the subsequent fixations (10, 11) focus on the correct headword *forma*. Next comes a fairly systematic skim over the sense indicators of the consecutive senses (fixations 12-19). Fixations 16 and 17 likely cover two one-line senses each, as the sense indicators lie within the area of foveal vision for these fixation points (at the viewing distance used in the experiment, foveal (= sharpest) vision covers a circle of approximately 2.5 centimetres in diameter, which in the vertical dimension corresponds to about three lines of text in our material).

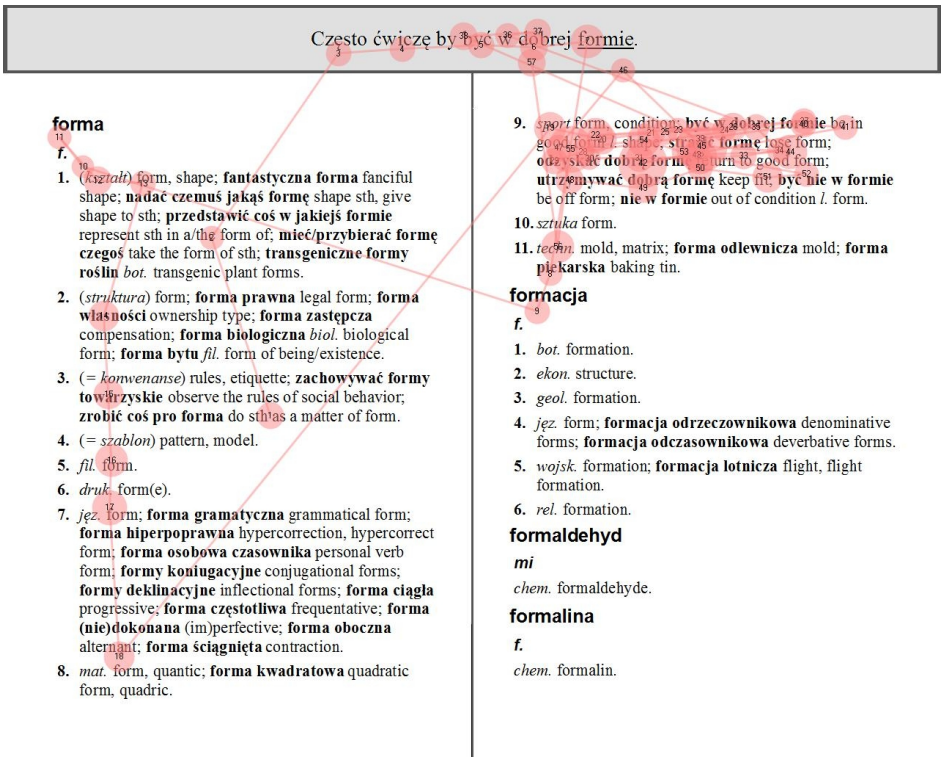

Figure 4: A typical scan path of a successful search.

Having reached sense 9 (our target sense), this high-proficiency participant spends a long time studying the sense, including the multi-word expressions. Much attention is directed (fixations 24-35) at the expression *być w dobrej formie* ('keep fit'), which is actually the expression used in the sentence cue. At this point, the participant makes a detour back to the second half of the sentence cue (fixations 36-38), apparently to check the match of the context against the expression located in sense 9. He then reconfirms the match (39-41), but goes on to read the remaining phrases nevertheless (fixations 42-55). A look at the two remaining senses (fixation 56) and return of the gaze to the sentence cue

conclude this successful search. We should stress again that about 80% of the searches recorded in the experiment were of this nature (though not necessarily going beyond the target sense).

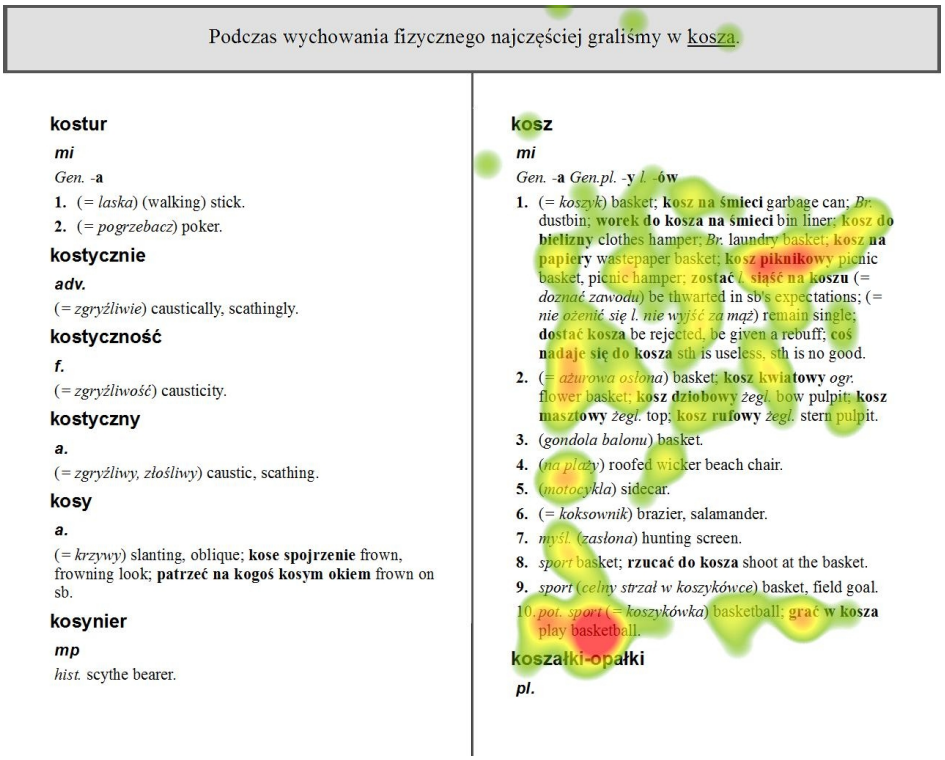

Figure 5: Fixation count heat map of a successful search for an entry-final sense by a low-proficiency participant. An animated version of this figure is available at http://vimeo.com/59560367.

Another example of a successful search, this time from a low-proficiency student, is given in **Figure 5**. The figure presents a so-called fixation count *heat map* (an animated version is available at http://vimeo.com/59560367). In this type of visualization, the more an area has been looked at, the hotter (redder) its colour (in non-colour print this appears as a darker shade). In this entry (*kosz*, 'basket'), our target was the very final sense (10). It represents an informal use referring to basketball as a game (*koszykówka* in general Polish). The fixation pattern indicates that the participant (LP07) reviewed the respective sense indicators and, having spent some time scanning (twice) sense one and examining some bold-type phrases in the earlier senses, correctly homed in on the final sense, and then identified a nested phrase which corresponds to the expression in the sentence cue, *grać w kosza* 'play basketball'. The challenge of this particu-

lar item lies in the fact that the entry includes, not just one, but three basketball-related senses, all of them marked with the domain label *sport*. Sense 8 refers to the circular net used in basketball, and sense 9 to a goal scored in the game.

However, participants did not necessarily examine all the senses. Just as suggested by Tono (1984), searches stopped at the first sense if the translation seemed to fit. This strategy was quite common among the low-proficiency participants, but only very occasional with high-proficiency students. **Figure 6** illustrates this difference in approach, using the entry *rakieta* ('missile') as an example. In this item, the sentence cue unambiguously referenced a Polish-made anti-aircraft missile.

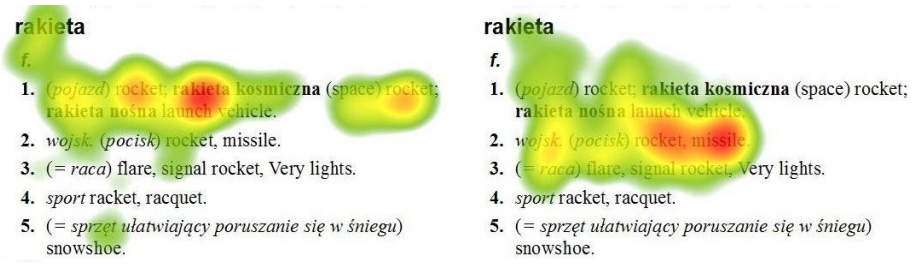

Figure 6: Two cumulative fixation heat maps for the item *rakieta* ('missile'), with the military sense number 2 being the target. On the left, unsuccessful consultations by low-proficiency students; on the right, successful consultations by high-proficiency participants.

The heat map on the left represents cumulative data from the five low-proficiency participants, four of whom opted for the incorrect first sense. On the right, data from high-proficiency participants are visualized: here, four out of five participants selected sense number 2, which best reflects the use of the word in the original sentence cue. Sense 1 is the most general sense and is indicated by the hyponym *pojazd* ('vehicle'). Sense 2 is the specific military sense and has two sense indicators: the abbreviated domain label *wojsk.* for *wojskowość* ('military'), and the near-synonym *pocisk* ('projectile'). The advanced learners were for the most part able to locate this specific military meaning, and the heat map reveals that they had studied the sense indicators. The lower-proficiency participants selected sense 1, which, admittedly, is not a completely wrong-headed choice in this case. Another interesting finding showing up in the heat map is that the first word of the phrase **rakieta kosmiczna** attracted a lot of attention from low-proficiency participants. The likely reason for this is the typography: it was the first bold-type element within this entry. This observation underscores the important role that typography plays in dictionary entries.

The above analysis illustrates an important tendency: while lower proficiency participants were often happy to skip any further senses once they were

reasonably satisfied they had identified the correct choice, participants in the high-proficiency group often preferred to continue looking at the remaining senses as well. This systematic strategy of high-proficiency participants resulted in success in the above case, but in many longer entries it led them to examine the remaining senses even after the correct earlier sense had already been positively identified, possibly wasting some time.

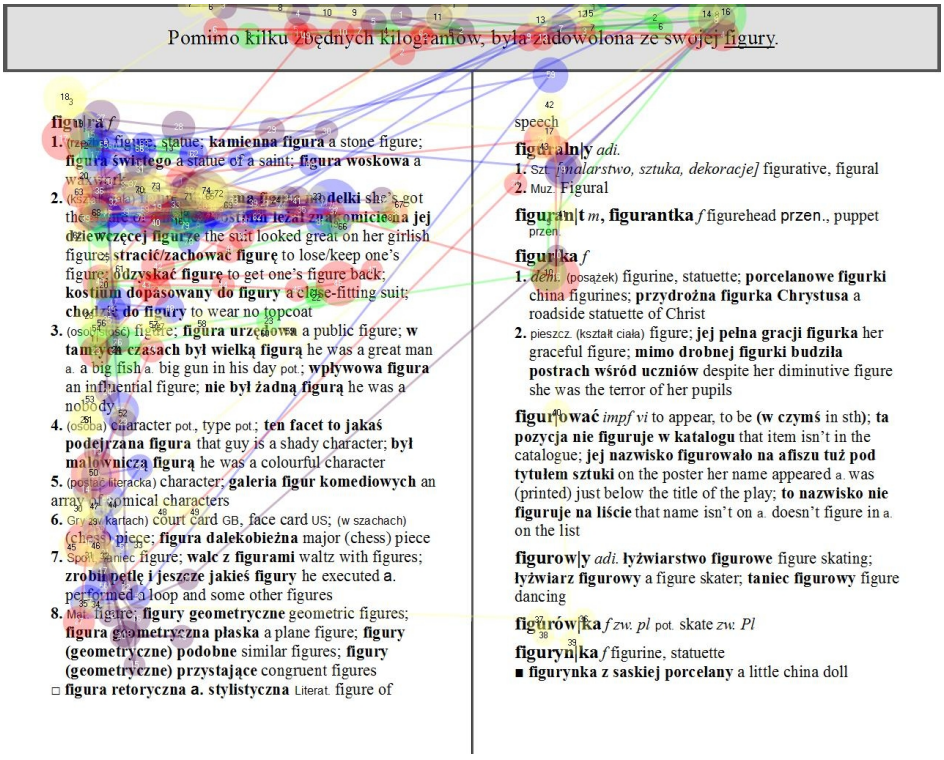

Figure 7: Cumulative gaze plot of the five HP participants consulting the entry for *figura*. Most looked at all senses, with sense 2 being the appropriate one.

A typical scenario illustrating this strategy of high-proficiency students is illustrated in **Figure 7**, a cumulative gaze plot of the five high-proficiency participants working with the entry for *figura* ('figure, shape'). The plot indicates that four out of five HP subjects examined all eight senses, and one participant viewed sense 3. Thus, all high-proficiency participants looked beyond the target sense 2. Most subsequently went back to the appropriate sense, but having checked the remaining senses first.

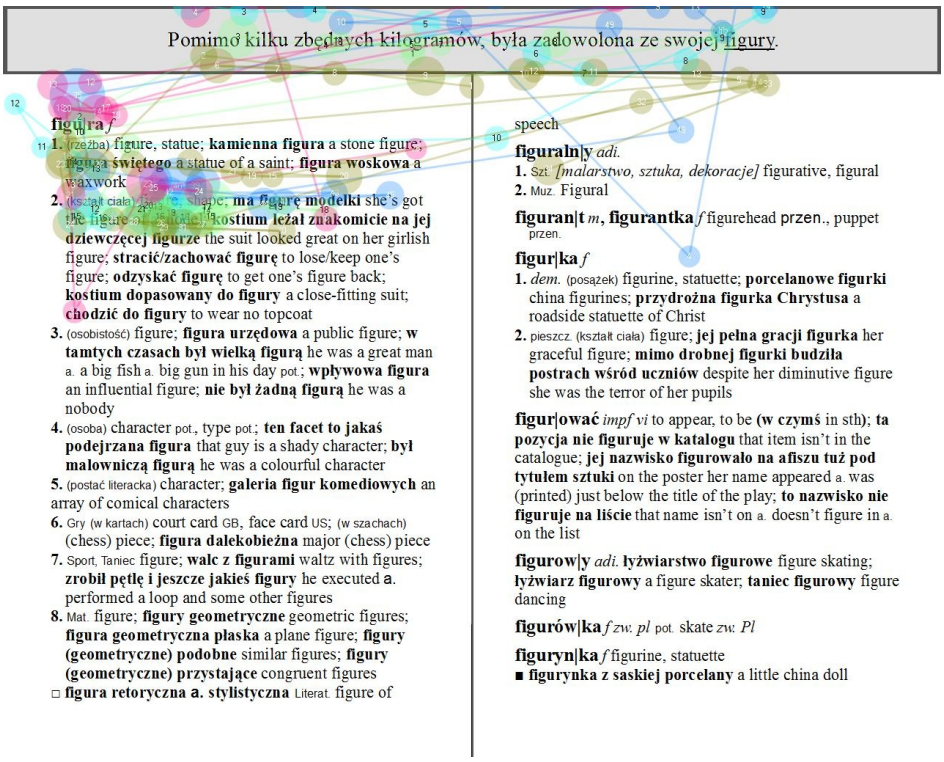

Figure 8: Cumulative gaze plot of the five LP participants consulting the entry for *figura*. All students stopped at the appropriate sense 2.

In such cases, the dominant strategy of our low-proficiency participants may be described as more economical than the one evident in the HP group: they dispensed with examining the remaining senses beyond the sense found relevant to the context of the sentence cue, in this case sense 2. In other words, once they were satisfied with sense 2 being the appropriate one, they moved on. This difference of strategy is evident in many items. On the other hand, low-proficiency students did not work any faster than the high-proficiency group (see section 5.1 above). However, it is possible that they worked more slowly in general and, had they decided to also examine further senses, they would have needed more time than the high-proficiency participants.

An interesting case of a relatively short entry where a substantial proportion of users have missed a specific equivalent is that of *promień* ('radius', as of a circle). Here, the wrong equivalent *ray* was supplied in four instances of unsuccessful searches (**Figure 9**). The required equivalent was *radius*, as used in geometry. This sense was clearly marked in the entry under sense 3, which held the domain label *geom.*, transparent to a Polish speaker. However, the four unsuc-

cessful users did not get to this sense. Instead, they registered the many instances of *ray* as an equivalent, and presumably concluded from the repeated tokens of *ray* that this equivalent is universal enough and it will do in this case as well.

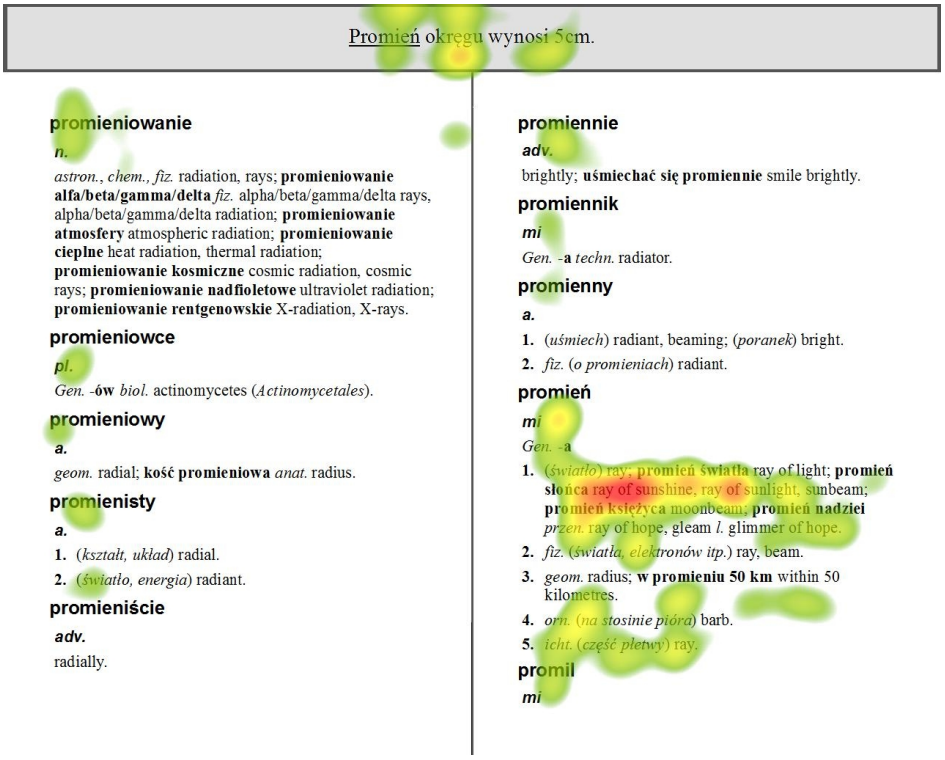

Figure 9: Cumulative fixation count heat map for the four students who gave the wrong equivalent of *promień* ('radius').

A very similar scenario was noted in the case of the item *poślizg* ('delay'). This would suggest that visual prominence of an equivalent which occurs several times within an entry can sometimes override the significance of sense indicators, effectively preventing dictionary users from selecting an isolated equivalent when exposure to several tokens of a more popular equivalent gives them false confidence that the latter will fit in just about any context. It is hard to come up with an acceptable lexicographic solution to this kind of problem. Giving extra salience to such isolated equivalents through typography might be one option, but it is controversial as it might in turn result in an overuse of this particular equivalent. Another avenue is to try to avoid repeating a frequent equivalent too many times, but rather to group together senses served by the same equivalent. Considered from this point of view, an equivalent-struc-

ture of the bilingual entry might be better than a source-language structure (Jarošová 2000; Adamska-Sałaciak 2006; Lew 2013).

5.6 Headword scan patterns

Our study focused on the inner access search process, i.e. the search within the entry once it has been identified, with a view to locating the relevant sense. However, since we used page-sized dictionary mock-ups, our data also allows some conclusions with regard to how dictionary users scan the dictionary page in search of the relevant headword. These findings, based mostly on raw (unfiltered) data, will be presented briefly, as we had not planned to investigate this aspect. However, we do want to include them, as they have important methodological implications for the design of similar studies in the future.

Gaze data indicate that our dictionary users typically scanned the headwords in the top-down direction. However, they did not necessarily start their search in the left-hand column (our mock-ups used the usual two-column layout). Instead, they often went from the end of the sentence cue to the top of the right-hand column, presumably because it was closer to the end of the sentence. Moreover, once they had reached the bottom headword in the column without finding the lemma sign sought, they would often cross over to the other column on the same level and then proceed in the upwards direction. These findings suggest that dictionary users, when scanning the headwords on a page, optimize their gaze path so it is as short as possible, even if this goes against the alphabetic sequence.

Another interesting finding was that, whereas scanning for information within the entry resembled patterns found in silent text reading, headword scans patterned differently. They tended to be more rapid and the landing point would often miss the headword, with an ensuing corrective movement, producing loose clusters of gaze points around some lemma signs. Some such clusters went undetected when a default setting of fixation detection filters was used. With the standard ClearView filter, a larger fixation radius (over 60 pixels) had to be used, combined with a low threshold for minimum fixation duration (40 milliseconds). This finding might explain why in Tono's (2011) study a few searches appeared to jump straight to the relevant section, without apparent evidence of the user scanning the page. It may be that the scan was rapid enough to have been filtered out, as a minimum fixation duration of 100 milliseconds was used. When a page is scanned in search of a relevant section, as in a dictionary headword scan, a shorter fixation duration setting should be used.

6. Summary and conclusion

The present study of sense look-up patterns with the use of an eye-tracking system has produced some interesting results. Overall, participants in both

groups performed fairly satisfactorily, having extracted the correct sense from a bilingual entry about four times out of five. As one would expect, many more erroneous choices were made in the lower-proficiency group. Eye-tracking data indicate that in many cases participants chose to examine every single sense in the relevant entry. High-proficiency participants tended to adopt this strategy even after they had identified the target sense; in contrast, low-proficiency participants usually terminated their search having found the relevant sense. Somewhat surprisingly, no evidence was found of shorter entries resulting in better success in selecting the relevant sense; in fact, a reverse tendency was noted. However, there was a weak (but not significant) correlation between the relative position of the target sense within the entry and success, with a tendency for earlier senses to be easier to identify. In addition, whenever the wrong sense was chosen, it was almost invariably one located higher up than the target sense, most commonly the first sense of the entry. This finding points to the special salience of the first sense of an entry, but no similar effect was found in our data for entry-*final* senses. To maximize success, lexicographers should try to place a translation equivalent with the broadest possible range of application in the first sense of a bilingual dictionary entry.

Our data also indicate that elements in bold attract significant attention of the users, and that they tend to interpret repeated occurrence of an equivalent as evidence of its universal application. This, in principle, is a valid inference, but it does make it more likely for users to ignore or miss the less common equivalents. To avoid this effect, the same equivalent should not be repeated too many times, if only there is a way to group lexicographic data so as to avoid such repetition.

An important finding of this study was that sense guiding elements occupy a significant proportion of the users' attention (between one-fifth and a quarter in terms of both fixation counts and relative dwell time). This proportion was quite stable across participants of both proficiency levels, suggesting that sense indicators in bilingual entries do fulfil the purpose for which they were designed.

This study has also tested the application of eye-movement tracking technology to the investigation of dictionary look-up processes, being one of the first to apply this instrumental approach within dictionary user research. The results demonstrate that eye tracking is a highly appropriate technique, as it provides detailed, first-hand information on users' visual scan patterns, both at the outer and inner search stages. From these patterns we infer information on which elements users consulted, how long they dwelled on them, and in what particular sequence, as well as whether they revisited particular elements. This type of data has not been available by any other technique. Unfortunately, the accuracy of currently available physical-object eye-tracking technology is not good enough to capture fine detail of dictionary structure. One solution used here is to use screen-based page mock-ups, but this inevitably detracts from the naturalness of the dictionary use situation.

In order for the use of eye tracking to be maximally useful, certain stan-

dards are needed to guide the design of similar studies in the future. One finding from the present study is the mean duration of fixation during the process of entry consultation. This has turned out to be about 300 milliseconds, that is some 30% longer than for normal text reading by native speakers. As variation is substantial, the minimal fixation threshold of around 80 milliseconds seems a valid cut-off value for entry reading. However, to capture headword scanning on a dictionary page, a threshold as low as 40 milliseconds may be needed, and a higher setting (at least 60 pixels in our case, or about two degrees of visual angle) for the fixation radius.

Overall, eye-tracking technology proves to be a highly fitting and fruitful approach for examining what happens in dictionary consultation, and should be used more widely.

References

Dictionaries

NKFD: Fisiak, J., A. Adamska-Sałaciak and P. Gąsiorowski (Eds.). 2003. *Nowy Słownik Fundacji Kościuszkowskiej. The New Kosciuszko Foundation Dictionary***.** New York: The Kosciuszko Foundation.

PWNO: Usiekniewicz, J. (Ed.). 2002. *Wielki Słownik Angielsko-Polski, Polsko-Angielski PWN-Oxford***.** Warszawa: PWN.

Other literature

Adamska-Sałaciak, A. 2006. *Meaning and the Bilingual Dictionary: The Case of English and Polish.* Frankfurt am Main: Peter Lang.

Al-Besbasi, I. 1991. *Translation Processes and Dictionary Use*. Ph.D. Thesis. Exeter: University of Exeter.

Bogaards, P. 1998. Scanning Long Entries in Learner's Dictionaries. Fontenelle, T., P. Hiligsmann, A. Michiels, A. Moulin and S. Theissen (Eds.). 1998. *EURALEX '98 Actes/Proceedings*: 555-563. Liège: English and Dutch Departments, University of Liège.

Duchowski, A. 2007. *Eye Tracking Methodology. Theory and Practice.* London: Springer.

Gerganov, A. 2007. Appendix A: Eye Tracking Studies with Tobii 1750 — Recommended Settings and Tests.

Harvey, K. and D. Yuill. 1997. A Study of the Use of a Monolingual Pedagogical Dictionary by Learners of English Engaged in Writing. *Applied Linguistics* 18(3): 253-278.

Jarošová, A. 2000. Problems of Semantic Subdivisions in Bilingual Dictionary Entries. *International Journal of Lexicography* 13(1): 12-28.

Just, M.A. and P.A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review* 87(4): 329-354.

Kaneta, T. 2011. Folded or Unfolded: Eye-Tracking Analysis of L2 Learners' Reference Behavior with Different Types of Dictionary. Akasu, K. and S. Uchida (Eds.). 2011. *Asialex2011 Proceedings Lexicography: Theoretical and Practical Perspectives*: 219-224. Kyoto: Asian Association for Lexicography.

Lew, R. 2004. *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English.* Poznań: Motivex.

Lew, R. 2010. Users Take Shortcuts: Navigating Dictionary Entries. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress*: 1121-1132. Ljouwert: Afûk.

Lew, R. 2013. Identifying, Ordering and Defining Senses. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 284-302. London: Bloomsbury Publishing.

Lew, R. and A. Dziemianko. 2006. Non-Standard Dictionary Definitions: What They Cannot Tell Native Speakers of Polish. *Cadernos de Traduçao* 18: 275-294.

Lew, R. and J. Pajkowska. 2007. The Effect of Signposts on Access Speed and Lookup Task Success in Long and Short Entries. *Horizontes de Lingüística Aplicada* 6(2): 235-252.

Lew, R. and P. Tokarek. 2010. Entry Menus in Bilingual Electronic Dictionaries. Granger, S. and M. Paquot (Eds.). 2010. *eLexicography in the 21st Century: New Challenges, New Applications*: 193-202. Louvain-la-Neuve: Cahiers du CENTAL.

Lowry, R. 2001–2013. *Vassarstats. Website for Statistical Computing*. Accessed on http://www.vassarstats.net/.

Mackintosh, K. 1995. *An Empirical Study of Dictionary Use in Version*. M.A. Thesis. Ottawa: University of Ottawa.

Nesi, H. and K.H. Tan. 2011. The Effect of Menus and Signposting on the Speed and Accuracy of Sense Selection. *International Journal of Lexicography* 24(1): 79-96.

Olsson, P. 2007. *Real-Time and Offline Filters for Eye Tracking*. M.A. Thesis. Stockholm: Royal Institute of Technology.

Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin* 124(3): 372-422.

Rayner, K. 2009. Eye Movements and Attention in Reading, Scene Perception, and Visual Search. *The Quarterly Journal of Experimental Psychology* 62(8): 1457-1506.

Simonsen, H.K. 2009a. Se — og Du Skal Finde: En Eyetracking-Undersøgelse Med Særlig Fokus på De Leksikografiske Funktioner. *Nordiske Studier i Leksikografi 11. Rapport Fra Konference Om Leksikografi i Norden. Finland 3.-5. Juni 2009*: 274-288. Tampere: Nordisk forening for leksikografi.

Simonsen, H.K. 2009b. Vertical or Horizontal? That Is the Question: An Eye-Track Study of Data Presentation in Internet Dictionaries. Paper presented at the Eye-to-IT Conference on Translation Processes, Sentence Processing and the Bilingual Mental Lexicon, Business School, Copenhagen, Denmark, April 25–29, 2009.

Simonsen, H.K. 2011. User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. *Hermes* 46: 75-101.

Tono, Y. 1984. *On the Dictionary User's Reference Skills*. B.Ed. Thesis. Tokyo: Tokyo Gakugei University.

Tono, Y. 1992. The Effect of Menus on EFL Learners' Look-up Processes. *Lexikos* 2: 230-253.

Tono, Y. 1997. Guide Word or Signpost? An Experimental Study on the Effect of Meaning Access Indexes in EFL Learners' Dictionaries. *English Studies* 28: 55-77.

Tono, Y. 2001. *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension.* Lexicographica Series Maior 106. Tübingen: Niemeyer.

Tono, Y. 2011. Application of Eye-Tracking in EFL Learners' Dictionary Look-up Process Research. *International Journal of Lexicography* 24(1): 124-153.

Wingate, U. 2002. *The Effectiveness of Different Learner Dictionaries. An Investigation into the Use of Dictionaries for Reading Comprehension by Intermediate Learners of German.* Lexicographica Series Maior 112. Tübingen: Niemeyer.